9.1 GLOSSARY

This glossary is divided into two parts: A listing of imputation terms as used by the StEPS imputation module and a listing of files used by StEPS to perform imputation processing.

9.1.1 Imputation Terms

(StEPS procedures and acronyms are in capital letters. Acronyms are spelled out in bold at the beginning of the definition. General terms are caps and lower case.)

ATREND One of the ratio item methods used in General Imputation. For ATREND, the user defines three auxiliary variables. The imputed value will be the value of the first auxiliary variable multiplied by the ratio of the second auxiliary variable to the third auxiliary variable. For example, for the variable v': $v' = A_1 * (A_2/A_3)$ This method differs from the other two ratio methods (AUXRAT and RATIO) in that all three variable values are taken from the same ID for which the imputed value is being calculated. The other two methods draw values from the imputation base.

Auxiliary Variable Any variable in addition to the variable being imputed. In StEPS a General Imputation calculation may have from one to ten auxiliary variables. Throughout this glossary, an auxiliary variable is identified by the symbol A_x, where the ‘x’ indicates the number of the auxiliary variable. The first auxiliary variable is indicated by A_1, the second by A_2, etc.

AUXRAT One of the item imputation methods used in General Imputation. This more generalized version of the RATIO method allows the user to define each of the three auxiliary variables from the imputation base. Uses the same method as RATIO imputation, but may use different items or different version of data (i.e., edited v = edited x * (sum of weighted v/ sum of weighted x).

Balance Complex A combination of variables consisting of one “total” item and at least two or more “detail” items such that the sum of the detail items must equal the total item. For example, $y = x_1 + x_2 + x_3$. StEPS offers two types of balance complex imputation:

Single 1-d balance complex: one total and a set of two or more details that sum to the total, e.g., $y = x_1 + x_2$.

Multiple 1-d balance complex: a set of two or more single-1d complexes such that the details in each of the complexes add to the same total item. (For example, $y = x_1 + x_2$ and $y = x_3 + x_4 + x_5$.) A detail item is a member of one and only one sub-complex.

BYGIMP **General Imputation Bypass Flag**. A field on the stat period control file used to exclude IDs from General Imputation. If BYGIMP = ' '(blank), General Imputation may be performed for this ID. If BYGIMP = 'B', General Imputation will not be performed.

BYIMPB **General Imputation Base Bypass Flag.** A field on the stat period control file used to exclude IDs from the imputation base. If BYGIMP = ' ' (blank), this ID will be included in the imputation base. If BYGIMP = 'B', this ID will not be included.

CAT or Categorical Variable An item used to limit the number of values for which a ratio-of-identicals or a mean. median, or mode must be calculated. For example, if a NAICS code is entered as a "Category variable" on the item imputation specification screen, all only eligible records with a matching NAICS value will be used in the required computations. Used by the AUXRAT, MEAN, MEDIAN, MODE and RATIO methods.

CRE8IMPB **Create Imputation Base**. A StEPS process that creates the imputation base. The CRE8IMPB record holds a set of logical conditions and optional extra operations specified by the user to ensure that only wanted data is included in the imputation base. If all of the conditions are met, the ID is eligible for inclusion in the base; otherwise it is not. The conditions will be tested and extra operations performed only if Option 6 on the "Run Imputation Screen" is checked, which turns CRE8IMPB is "on." If status is not "on", the conditions will not be used.

DATREND A ratio item methods used in General Imputation. The user defines three auxiliary variables. The imputed value will be the value of the first auxiliary variable multiplied by the ratio of the second auxiliary variable to the third auxiliary variable. For example, for the variable v': v' = A_1 * (A_2/A_3) The first variable comes from the same observation as the item being imputed. The second and third variables come from a separate data set specified by the user. (See Donor Imputation.)

Donor An ID record, usually an external data set, from which data items may be used to impute one or more data items for a recipient.

Donor Imputation The use of data from one or more donor ID records to provide item data for a recipient ID record when the recipient ID's data are missing or determine to be invalid.

DVALUE A logical and direct substitution item imputation method used in General Imputation. Differs from the VALUE method in that the auxiliary variable used to provide a value for the imputed item is from a donor ID, usually provided by an external data set specified by the user.

General Imputation StEPS module that imputes data using estimator type techniques and adjusts balance complexes so that detail items sum to total items. Values changes are flagged as imputed data.

GIBS **General Imputation Balance Complex Specifications** file. File that contains the list of detail items necessary to perform general balance complex imputation. StEPS refers to this file throughout the imputation process.

GIMP **General Imputation Parameter File**. File that contains information about the balance complexes that exist for a given survey. This file will contain a record for each balance complex in the survey. Each record specifies the items making up the associated balance complex and defines actions to take for each of the corresponding out-of-balance and missing-data conditions that could occur within the complex.

GIS **General Imputation Specifications file**. File that contains the formulas, equations, method orders, categories, etc. necessary to perform general item imputation. StEPS refers to this file throughout the imputation process.

GISGLBL The GIS information plus information needed for a specified balance complex imputation (TESTNUM, YITEM, NUMDET, and XITEM1-XITEM10).

IMPREJ **Imputation Reject File**. File built by StEPS that contains all the ID/Item record combinations for which values will be imputed. StEPS looks at edit rejects, flags, and a variety of other criteria to determine which cases should be included. This file contains a record for each test failure and for total and detail items that are out of balance.

IMPREJUD The unduplicated version of the Imputation Reject File. This file contains all the ID/Item record combinations that failed balance complexes and edit tests.

IMPACT **Imputation Action Flag**. A field on the item file that controls whether an item value will be imputed. Values include:

blank = Impute the item and **exclude** from base if the item failed any of the imputation tests. Do not impute the item and **include** in base if the item passed all of the imputation tests.

Y = Impute regardless of whether or not the item failed any of the imputation tests and do not include in base.

N = Do not impute the item regardless of whether or not the item failed any of the imputation tests and do not include in base.

X = Exclude this item from the imputation base but impute the item as if IMPACT flag was blank.

IMPFLG **Imputation Flag**. A flag set by the imputation program to identify the imputation routine that created the data. If DATAFLAG is changed by the user, this field will be blanked out. Values are: A = SUM; B = Custom Method 1; C = Custom Method 2; D = RESIDUAL; E = RESIDUA; F = Free Form; G = SIMPREG; H = MULTREG; I = RATIO IMPALL; J = AUXRAT; K = RAKEIMP; M = MEAN; N = Unable to impute; O = ROUND; P = PCTRAT.

Imputation Base A file containing values drawn from the survey database for use in imputation calculations.. This file is created during imputation processing. For example, if an item is to be imputed via the MEAN method (see below), the values used to produce the mean will be drawn from the imputation base.

MEAN A method for imputing an item's value by inserting the mean value, as calculated in the imputation base, of all the reported values for an auxiliary variable. This method may or may not use a category variable. The auxiliary variable used to perform the calculation could be the same item as the one being imputed, the same item from a different statistical period, or a different item from the same statistical period. For example, if imputing for the variable v': $v' = \bar{A}_1$ **.**

MEDIAN Imputing an item's value by inserting the median value, as calculated in the imputation base, of all the reported values for an auxiliary variable. This method may or may not use a category variable. The auxiliary variable used could be the same item as the one being imputed, the same item from a different statistical period, or a different item from the same statistical period.

MODE Imputing an item's value by computing the most frequently occurring value of an auxiliary variable and assigning this mode value to the item to be imputed. The auxiliary variable could be the same item as the one being imputed, the same item from a different statistical period, or a different item from the same statistical period.

NSK **Not Specified by Kind**. A balance complex method. If the user specifies a variable name as a NSK variable, the value of that variable will be set to the residual from the balance complex computation ($R = y- \sum x$). If the user specifies the NSK variable as "MAX," the value of the residual will be added to the detail item with the largest value. If the user specifies the NSK variable as one of the detail items, the residual will be added to that detail item.

PRODUCT An item imputation method. The imputed value is the product of two or more other values. For example, if imputing for the variable v’: $v' = A_1 * A_2$.

RAKE A balance complex method available in both Simple Imputation and General Imputation for adjusting balance complexes when data is missing for more than one detail item. The Simple Imputation version can be used when the total and all of the detail items are non-negative. The General Imputation version can handle negative detail items. The RAKE option, first computes the residual ($R = y- \sum_{nm}$) where $\sum_{nm}$ is the sum of the non-missing detail items. It then divides the value of the residual by the number of items missing data and assigns this value to each of the items will missing values. (The formula is $x_i^{/}= x_i(y/\sum_{nm} x_j) = x_i(1+R/\sum_{nm} x_j)$,

RAKEIMP Action available in General Imputation for adjusting balance complexes when data is missing for more than one detail item and when the balance complex contains imputed data. If a detail item has an imputation flag, this method replaces that value with a raked value that balances the complex.

RATIO An item imputation methods used in General Imputation. It is the same as AUXRAT except that the user must define only one auxiliary variable from the imputation base. This method first produces a *ratio of identicals*. (See definition below.) The numerator in the ratio of identicals is the item to be imputed, and the denominator is the first auxiliary variable, which may be a different item, a different data version of the same item, or data for that item from a different stat period.

Ratio of Identicals A ratio produced from two variables within the data base, usually sorted by some categorical variable. One example might be to compare current year sales values to prior year sales values by NAICS (i.e., $r = \text{Sales}_{\text{current}}/\text{Sales}_{\text{prior}}$). Other examples include expenses to sales ratios by NAICS or inventory to sales ratios by NAICS, with all comparisons made within the same statistical period (i.e, r = current expenses/current sales or current inventory/current sales). A ratio of identicals may also be referred to as an Industry Average or Industry Trend. Ratios of identicals are used in the RATIO and AUXRATIO methods.

Recipient An ID record for which one or more data items require imputation. In door imputation, the ID the receives data from the donor(s).

RECODE In General Imputation, the name of a macro that creates a temporary variable on the GFAT file, usually used to create category (CAT) variables that define imputation cells, objects to be imputed, auxiliary variables, and variables in imputation conditions. Recodes are run when the imputation fat file is created and are not based on any item values within imputation.

RESIDUA One of the logical and direct substitution item imputation methods used in General Imputation. The imputed value is computed for an auxiliary variable by subtracting the sum of a series of values from a reported total value. For example, if imputing for the variable v’: $v' = A_1 - (A_2 + A_3 + A_4)$.

RESIDUAL A single 1-dimensional imputation method used if only one detail in a balance complex is missing. It assigns the missing detail the value of the difference between the total and the sum of the nonmissing details. For example, if only the value for 3rd quarter payroll is missing, assign the value equal to the reported total minus the sum of 1st quarter, 2nd quarter, and 4th quarter payroll (e.g., $x_3 = y - (x_1 + x_2 + x_4)$.

ROUND Action for adjusting balance complexes. Divides details by 1000, then rakes (x_i /1000 replaces x_i in rake formula).

Simple Imputation The StEPS imputation module that imputes data values considered to be equivalent to reported data. Items changed are flagged as reported.

SIMPREG The single regression method. The imputed value is the result of multiplying the regression coefficient by a single auxiliary variable. For example, if imputing for the variable v’: v’ = ß$_{v,A}$(A_1). This is also written $v' = \beta_1 z_1$.

SUM One of the four logical and direct substitution item imputation methods used in General Imputation. The imputed value will be the sum of two or more auxiliary variables. For example, if imputing for the variable v’: $v' = A_1 + A_2 + A_3$

Absolute Tolerance The largest permissible value of the deviation of the sum of the details from the total in order to use the imputation method. If an out-of-balance complex exceeds a specified absolute tolerance (expressed as an integer), StEPS will not attempt to balance the complex. If no tolerance is specified, balancing will always be attempted.

Relative Tolerance The largest permissible value of the deviation of the sum of the details from the total divided by the total in order to use the imputation method. A relative tolerance is met if the absolute value of the ratio of the residual to the total is less than the relative tolerance. If an out-of-balance complex exceeds the survey-specified relative tolerance, balancing will not be attempted. If no tolerance is specified, balancing will always be attempted.

VALUE One of the logical and direct substitution item imputation methods used in General Imputation. The imputed value is simply the value from another data item, such as producing a value for third quarter payroll by directly inserting the reported value for second quarter payroll. For example, if imputing for the variable v': v' = A_1

YSUMX A balance complex adjustment method. The total is set to the sum of the non-missing detail items. For example, set y to $\sum_{nm} x_i$, where $\sum_{nm} x_i$ is the sum of the non-missing x_i.

ZERO_SET An action to adjust balance complexes. The missing detail item(s) are set to zero.

9.1.2 Imputation Files

StEPS creates the following files during imputation processing.

USER-CREATED FILES

General Imputation

- GIS: General Imputation Specifications file.
- GIBS: General Imputation Balance Complex Specifications file.

Simple Imputation

- SIMP1: Specifications for Simple Imputation module free-form imputation. Used to generate imputation code.
- SIMP2: Specifications for Simple Imputation module balance complex imputation. Used to generate imputation code.

PROCESSING FILES - General Imputation

- GFAT: FAT record–created specifically for imputation–contains all IDs, all items, all versions of the data and all flags for all imputation relative stat periods.
- IMPDETL: A report of all non-missing imputation results.

- IMPREJ: Although called the imputation reject file, it is not a list of records that have failed imputation. Rather, it holds the list of records from the GFAT which have failed edit tests and which must be imputed..

- IMPREJUD: The unduplicated form of IMPREJ.

- Imputation Base files

  - BASEDETL: Individual items in imputation base
  - BASESUM: Sum by category variable and specification of item values